PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort profile: NeuroBlu, an Electronic Health Record Trusted Research Environment to support mental healthcare analytics with real-world data
AUTHORS	Patel, Rashmi; Wee, Soon Nan; Ramaswamy, Rajagopalan; Thadani, Simran; Tandi, Jesisca; Garg, Ruchir; Calvanese, Nathan; Valko, Matthew; Rush, A.; Rentería, Miguel; Sarkar, Joydeep; Kollins, Scott

VERSION 1 – REVIEW

REVIEWER	Shiner, Brian Geisel Sch Med Dartmouth
REVIEW RETURNED	14-Oct-2021

GENERAL COMMENTS	Thank you for the opportunity to review this manuscript. One quick note to start was that there was an appendix that was almost 90 pages long associated with this manuscript, which I only skimmed to check for protected health information given the topic area. There were some counts below 11 on page 93 for Armed Forces Europe and Armed Forces Pacific. While I do not think any individuals would be identifiable based on these counts, the authors may want to check with their parent organizations to see whether these figures should be suppressed in publication. The overall point of this manuscript appears to be describing an electronic registry of mental health care data. The registry has already served as the data source for several studies. If published, this manuscript could serve as a resource for reviewers and readers attempting to interpret the results of future studies from the registry or guide researchers in considering whether the registry has the information that they need. To me, the major weakness of this registry is that there is no patient-reported outcome data. There is a brief allusion to this in the introduction, but no mention in the “strengths and limitations” section. Overall, this manuscript is of limited general interest, but would be valuable inclusion in the literature.

REVIEWER	Resnick, Philip University of Maryland at College Park
REVIEW RETURNED	16-Nov-2021

GENERAL COMMENTS	The authors present a description of NeuroBlu, a commercially available infrastructure for secure analysis of and machine learning with electronic health records in the study of mental health. The

framework is currently populated with records from more than a half million patients, and according to the website it is expected to grow to 2 million patients by the end of 2021.

In their collaboration section, the authors encourage collaborations, including with non-commercial participants like academic institutions. However, no details are provided on the web site on how such engagement works, nor its cost. More generally, I can see no evidence that this infrastructure is available, or will ever be made widely available, for academic or non-profit collaborations, except on unspecified commercial terms. A web search turns up an August 2021 call for proposals that would provide a cash grant of 3000 Euros plus six months of access to the platform, conditioned on permitting use of participant names and logos for apparent marketing purposes. Based on the information presented, therefore, the submitted article is a hybrid: it is both a genuine contribution to the research literature, and it could be interpreted as promoting a commercial product that may or may not be relevant for researchers with whom the enterprise is not commercially engaged. I should emphasize that all of this is completely above-board: connections with the company marketing NeuroBlu are made clear in the author affiliations and in the competing interests statement. That said, the appropriateness of a publication in this journal with those competing interests is a question of community norms and editorial policy and decision-making. As a reviewer, I am proceeding by leaving that aside and commenting on the paper under the assumption that there are no conflict/competing interest concerns.

To begin, there is an enormous need for data and infrastructure of the kind described here. As the explosion of work in machine learning over the last decade has made clear, the most effective path to progress is for entire communities to work on related problems in shared datasets, at scale. From a methodological perspective, healthcare research is severely hampered by privacy and data security concerns -- often driven as much by the fears of potential data provider organizations as by any actual laws or regulations. Secure data enclaves of the kind described here are in my opinion the best solution to this problem: by bringing researchers to the data, rather than vice versa, it is possible to mitigate the risks while still enabling community-level access. The research environment and patient cohort described here are constructed carefully and thoughtfully, managing to preserve a wealth of valuable data with sufficient size and content to support an enormous range of analyses. The application of natural language processing (NLP) methods to extract additional structured data fields from unstructured portions of the record is an important advance over data enclave approaches that support analysis only for structured data fields, and the R and Python capabilities that are sketched out in the paper also would appear to constitute an important increase in flexibility over the capabilities of traditional statistics packages.

With that said, I do have some comments, including a number of ways that the presentation in the paper can be improved.

First, it would be valuable for the authors to situate this particular platform and mental health dataset within the broader literature on sharing data for analysis of clinical records. In terms of datasets that are disseminated out to researchers, a relevant comparison here would be with MIMIC-III, which is as far as I know the only widely available EHR dataset of any appreciable size. In terms of closely related warehouse/enclave efforts bringing researchers to clinical data, there are a number of existing approaches out there, some examples including NORC's Health Data Enclave, the Coleridge Initiative's Administrative Data Research Facility (ADRF), Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records (CREATE, Liu, S. et al. 2020, JMIR medical informatics, 8(10), e17376), and others. The Observational Medical Outcomes Partnership/Observational Health Data Sciences and Informatics (OMOP/OHDSI) effort came to mind also; this is discussed in the supplemental materials but I think it would fit in the kind of foregrounded discussion I am suggesting.

Second, because the extraction of discrete features from free text in the record is part of the central contribution here, it is important to situate that aspect of the work a bit more fully in the context of relevant prior work on NLP for clinical records, as opposed simply to referring to Mukherjee et al. 2020/2021. (Also, the one-sentence summary of NLP on page 7, line 39 is strikingly inadequate in terms of conveying what NLP actually is for readers unfamiliar with it!) For example, the BioNLP special interest group of the Association for Computational Linguistics has long run a series of workshops that include work on that topic, and ACL has also hosted multiple workshops on clinical NLP (https://aclanthology.org/venues/clinicalnlp/); in addition, there have been multiple i2b2 NLP challenges (now re-branded as n2c2), including tasks on de-identification and on clinical information extraction. Specifically with regard to NLP analysis for mental health, there may also be relevant discussion or references in community-level shared task efforts at the ACL Workshops on Computational Linguistics and Clinical Psychology (clpsych.org, especially the 2021 shared task using NORC's secure data enclave) or the eRisk conferences (erisk.irlab.org). Obviously a comprehensive literature review on NLP for clinical records or even just mental health would be outside the scope of the article, but appropriate references (including pointers to suitable surveys) would go a long way toward addressing this issue.

Third, it is probably at least worth mentioning the large and growing body of work on machine learning and modeling for mental health using data *outside* EHRs -- analyzing data in what Coppersmith et al. (2018, Biomedical Informatics Insights, 10, 1178222618792860) call the "clinical whitespace" -- particularly in light of the focus on social variables. The comment about about references and surveys is applicable here, as well.

As a more substantive comment, I have concerns about presenting data that's constructed using NLP as if the output of algorithms were fully reliable. Simply saying that records are de-identified, for example, masks the fact that even good automated de-identification makes mistakes. If de-identification successfully identified every single piece of identifying information for all of a patient's records 99.9% of the time, which is extraordinarily optimistic, that would mean that in the anticipated 2M-patient dataset one should expect PHI to be retained in the dataset for 2000 participants. Similarly, the automatic extraction of things like discrete sociodemographic variables from free-text is great, but simply including binary variables as if they were any other variable masks the fact that the NLP models are producing non-trivial numbers of false negatives and false positives. The presence of an NLP-derived positive variable and the presence of a an ICD-10 code put in the record by a certified professional coder are two utterly different things. It's certainly true that even human annotations can have errors (and therefore one should evaluate chance-corrected inter-annotator reliability for human annotation tasks), but automated performance is typically well below that upper-bound level of performance.

In the absence of 100% human validation of the NLP results, therefore, in my opinion it's quite important to include an expected level of (in)correct classification along with automatically constructed variables -- not just in a published article (e.g. the results aggregating into eight categories, supplement page 16, are a start), but visibly for users in the platform whenever analysis results are being reported. An even better approach would be to provide users with confidence estimates at the per-item level, i.e. a confidence associated with each of the 241 NLP-derived binary features in every data record. One could even imagine the UI giving users control over how confidently an NLP-derived variable must be assigned in order to treat it as present in the record. (Cf. Jiang et al. 2006, "How does the system know it’s right? Automated confidence assessment for compliant coding." In Perspect Health Inf Manag, Computer Assisted Coding Conference Proceedings.)

To be fair, the issue I'm raising here is present in many if not most user-facing NLP applications; for example, commercial sentiment platforms will frequently generate reports about percentages of positive, negative, and neutral statements in some dataset being analyzed as if one could be sure that those sentiment labels have all been assigned correctly. Still, I would like to think that clinical applications should be adhering to a higher standard, in much the same way that responsibly conducted political surveys report confidence intervals, rather than just giving point-estimate percentages.

Finally, I also think it's important to think about this work in the context of broader developments in the analysis of medical data. The bar has been significantly lowered in recent years for applying rich machine-learning approaches in a wide variety of domains, but, as noted above, progress in the clinical space has been significantly hampered by lack of access to the same large, shared datasets by diverse teams. NeuroBlu represents a solid step in the right direction (to the extent it is actually available to a rich variety of researchers who can afford to use it) in terms of scale plus the ability to bring features communicated via text to the surface for statistics/analytics. A significant remaining obstacle to further progress, however, is the fact that a single, proprietary NLP approach is baked into the platform. And, in fact, it is an approach (LSTM with static word2vec embeddings; see Mukherjee et al. 2020) that has already been outstripped by more recent developments (contextual embeddings and the pretraining/finetuning approach using transformers). That doesn't make what's being done here wrong or inadequte, it's more of a missed opportunity (so far): it would be quite interesting to consider whether it would be possible to extend the platform's R/Python Code Engine capabilities to include a fuller range of machine learning approaches, including direct access to the (de-identified) text within the secure environment, so that users of the platform

	could themselves explore new classification approaches as the state of the art improves, and also expand to more NLP-derived features as motivated by the particular questions they are investigating.

VERSION 1 – AUTHOR RESPONSE

Response to reviewers - bmjopen-2021-057227

Reviewer: 1
Dr. Brian Shiner, Geisel Sch Med Dartmouth
Comments to the Author:
Thank you for the opportunity to review this manuscript.

One quick note to start was that there was an appendix that was almost 90 pages long associated with this manuscript, which I only skimmed to check for protected health information given the topic area. There were some counts below 11 on page 93 for Armed Forces Europe and Armed Forces Pacific. While I do not think any individuals would be identifiable based on these counts, the authors may want to check with their parent organizations to see whether these figures should be suppressed in publication.

/*Thank you for highlighting this. We have updated the eTable 7 in the supplementary material to suppress any numerical figures less than 20.*/

The overall point of this manuscript appears to be describing an electronic registry of mental health care data. The registry has already served as the data source for several studies. If published, this manuscript could serve as a resource for reviewers and readers attempting to interpret the results of future studies from the registry or guide researchers in considering whether the registry has the information that they need. To me, the major weakness of this registry is that there is no patient-reported outcome data. There is a brief allusion to this in the introduction, but no mention in the “strengths and limitations” section.

Overall, this manuscript is of limited general interest, but would be valuable inclusion in the literature.

/*Thank you for your supportive comments. We agree that patient-reported outcome data could substantially enhance the utility of electronic health record (EHR)-derived real-world data and we have already commented on this in the *Future plans* section of the manuscript **(Page 16, Paragraph 1)**:

“We aim to incorporate these tools into the source EHR alongside patient-reported outcome data to provide clinicians with actionable insights to support real-time clinical decision making. This approach could help to improve clinical outcomes by better personalizing mental healthcare and reducing delays to effective treatment.”

We have further updated the *Strengths and limitations* section to discuss this further and highlight how patient-reported outcome data could address the limitations of EHR data to support real-world evidence (RWE) generation and clinical decision support **(Page 15, Paragraph 2)**:

“At present, the dataset only includes clinician-recorded EHR data and does not include any patient reported outcome measures (PROMs) such as patient-recorded symptom or medication side effect scales. PROMs collected in between periods of clinical contact could help to address the limitations of EHR data described previously by providing patient-rated data on mental health symptom burden and response to treatment that could support more nuanced evidence generation and clinical decision making.”*/

Reviewer: 2

Prof. Philip Resnick, University of Maryland at College Park
Comments to the Author:

The authors present a description of NeuroBlu, a commercially available infrastructure for secure analysis of and machine learning with electronic health records in the study of mental health. The framework is currently populated with records from more than a half million patients, and according to the website it is expected to grow to 2 million patients by the end of 2021.

In their collaboration section, the authors encourage collaborations, including with non-commercial participants like academic institutions. However, no details are provided on the web site on how such engagement works, nor its cost. More generally, I can see no evidence that this infrastructure is available, or will ever be made widely available, for academic or non-profit collaborations, except on unspecified commercial terms. A web search turns up an August 2021 call for proposals that would provide a cash grant of 3000 Euros plus six months of access to the platform, conditioned on permitting use of participant names and logos for apparent marketing purposes. Based on the information presented, therefore, the submitted article is a hybrid: it is both a genuine contribution to the research literature, and it could be interpreted as promoting a commercial product that may or may not be relevant for researchers with whom the enterprise is not commercially engaged. I should emphasize that all of this is completely above-board: connections with the company marketing NeuroBlu are made clear in the author affiliations and in the competing interests statement. That said, the appropriateness of a publication in this journal with those competing interests is a question of community norms and editorial policy and decision-making. As a reviewer, I am proceeding by leaving that aside and commenting on the paper under the assumption that there are no conflict/competing interest concerns.

/*Thank you for your comments on our manuscript. We present a cohort profile of a de-identified mental health EHR dataset that can be analyzed through NeuroBlu, a secure, trusted research environment (TRE) to generate RWE to better understand mental disorders and support the development and implementation of more effective treatments. We believe that TREs are crucial to broadening access to real-world datasets that have historically been restricted to individuals or organizations with local access to data, thus limiting their utility to generate RWE. We are committed to ensuring that NeuroBlu is available to a wide range of individuals and organizations ranging from the academic and non-profit to regulatory and commercial sectors. The grant that you cite (NeuroBlu Award 2021 - https://www.holmusk.com/news/neuroblu-award) is one example of initiatives we are undertaking to make the dataset available to academic and other small organizations. We are committed to working closely with academic and non-profit organizations to support mental health RWE generation and have updated the *Collaborations* section of the manuscript **(Page 16, Paragraph 2)** to further describe our efforts in this area:

"We actively collaborate with a academic partners and mental healthcare providers to support mental health RWE generation.[62]"*/

To begin, there is an enormous need for data and infrastructure of the kind described here. As the explosion of work in machine learning over the last decade has made clear, the most effective path to progress is for entire communities to work on related problems in shared datasets, at scale. From a methodological perspective, healthcare research is severely hampered by privacy and data security concerns -- often driven as much by the fears of potential data provider organizations as by any actual laws or regulations. Secure data enclaves of the kind described here are in my opinion the best solution to this problem: by bringing researchers to the data, rather than vice versa, it is possible to mitigate the risks while still enabling community-level access. The research environment and patient cohort described here are constructed carefully and thoughtfully, managing to preserve a wealth of valuable data with sufficient size and content to support an enormous range of analyses. The application of natural language processing (NLP) methods to extract additional structured data fields from unstructured portions of the record is an important advance over data enclave approaches that support analysis only for structured data fields, and the R and Python capabilities that are sketched out in the paper also would appear to constitute an important increase in flexibility over the capabilities of traditional statistics packages.

/*Thank you for your supportive comments and we agree with the pressing need to improve availability and access to real-world datasets, particularly in mental healthcare.*/

With that said, I do have some comments, including a number of ways that the presentation in the paper can be improved.

First, it would be valuable for the authors to situate this particular platform and mental health dataset within the broader literature on sharing data for analysis of clinical records. In terms of datasets that are disseminated out to researchers, a relevant comparison here would be with MIMIC-III, which is as far as I know the only widely available EHR dataset of any appreciable size. In terms of closely related warehouse/enclave efforts bringing researchers to clinical data, there are a number of existing approaches out there, some examples including NORC's Health Data Enclave, the Coleridge Initiative's Administrative Data Research Facility (ADRF), Cohort Retrieval Enhanced by Analysis of Text from Electronic Health Records (CREATE, Liu, S. et al. 2020, JMIR medical informatics, 8(10), e17376), and others. The Observational Medical Outcomes Partnership/Observational Health Data Sciences and Informatics (OMOP/OHDSI) effort came to mind also; this is discussed in the supplemental materials but I think it would fit in the kind of foregrounded discussion I am suggesting.

/*We have updated the Introduction section **(Page 4, Paragraph 3)** to provide further information on other initiatives to improve access to EHR datasets such as those you have cited (e.g., through secure assembly/query/analytics frameworks or common data models):

"EHR datasets have previously supported RWE generation in critical care[13] and infrastructure based on common data models[14] has been developed to enable researchers to securely analyze RWD.[15,16]"*/

Second, because the extraction of discrete features from free text in the record is part of the central contribution here, it is important to situate that aspect of the work a bit more fully in the context of relevant prior work on NLP for clinical records, as opposed simply to referring to Mukherjee et al. 2020/2021. (Also, the one-sentence summary of NLP on page 7, line 39 is strikingly inadequate in terms of conveying what NLP actually is for readers unfamiliar with it!) For example, the BioNLP special interest group of the Association for Computational Linguistics has long run a series of workshops that include work on that topic, and ACL has also hosted multiple workshops on clinical NLP (https://aclanthology.org/venues/clinicalnlp/); in addition, there have been multiple i2b2 NLP challenges (now re-branded as n2c2), including tasks on de-identification and on clinical information extraction. Specifically with regard to NLP analysis for mental health, there may also be relevant discussion or references in community-level shared task efforts at the ACL Workshops on Computational Linguistics and Clinical Psychology (clpsych.org, especially the 2021 shared task using NORC's secure data enclave) or the eRisk conferences (erisk.irlab.org). Obviously a comprehensive literature review on NLP for clinical records or even just mental health would be outside the scope of the article, but appropriate references (including pointers to suitable surveys) would go a long way toward addressing this issue.

/*We agree that NLP is a key component of our approach to enhance the utility of mental health EHR data for RWE generation. Our article is aimed at a medical readership in accordance with the aims and scope of the BMJ Open and, as you point out, a technical review of the application of NLP to EHR data would be beyond the scope of this manuscript. However, we have added additional information to the section on *NLP pipeline to extract MSE and social history data* **(Page 7, Paragraph 5)** to describe the application of NLP specific to mental health EHR data for a clinical readership:

"NLP is a text mining technique that enables automated extraction and classification of features from unstructured free text that would be otherwise unfeasible to manually extract by reading through large volumes of text. The application of NLP to mental health EHR data involves a series of processes to develop algorithms that can identify clinically meaningful concepts for secondary analysis. These processes include: (i) Data assembly: identifying a collection of relevant documents (the corpus), (ii) Annotation: Clinical experts annotate a selection of the corpus to classify meaningful features to generate a training set to develop the NLP model and a reference set to evaluate its performance, (iii) Preprocessing: preparation of the corpus for NLP model development including stop word removal, stemming, lemmatization and parts of speech tagging, (iv) Featurization: classifying text within the corpus into different features (e.g. parts of speech, word vectors or embeddings, sentiment or temporal features) and (v) Analysis: development of NLP algorithm using a rules-based or machine learning approach using the training set. The resulting models are evaluated on the previously annotated reference set and tuned to maximize accuracy as measured through precision (positive predictive value), recall (sensitivity) and F1 measure (harmonic mean of precision and recall).[37] If a sufficiently accurate models are developed, they can be applied to the entire corpus to generate structured data on clinically meaningful features of interest[27] to support mental health RWE generation.[28,38–42] NLP models have also been investigated as a potential method to screen social media data to identify risk of suicide.[43]"*/

Third, it is probably at least worth mentioning the large and growing body of work on machine learning and modeling for mental health using data *outside* EHRs -- analyzing data in what Coppersmith et al. (2018, Biomedical Informatics Insights, 10, 1178222618792860) call the "clinical whitespace" -- particularly in light of the focus on social variables. The comment about about references and surveys is applicable here, as well.

/*Thank you for highlighting this important work. We agree that the application of NLP to social media data has the potential to improve our understanding of the associations of social factors with poor mental health outcomes and have cited it in the *Data Pipeline* section **(Page 7, Paragraph 5)**.*/

As a more substantive comment, I have concerns about presenting data that's constructed using NLP as if the output of algorithms were fully reliable. Simply saying that records are de-identified, for example, masks the fact that even good automated de-identification makes mistakes. If de-identification successfully identified every single piece of identifying information for all of a patient's records 99.9% of the time, which is extraordinarily optimistic, that would mean that in the anticipated 2M-patient dataset one should expect PHI to be retained in the dataset for 2000 participants. Similarly, the automatic extraction of things like discrete sociodemographic variables from free-text is great, but simply including binary variables as if they were any other variable masks the fact that the NLP models are producing non-trivial numbers of false negatives and false positives. The presence of an NLP-derived positive variable and the presence of a an ICD-10 code put in the record by a certified professional coder are two utterly different things. It's certainly true that even human annotations can have errors (and therefore one should evaluate chance-corrected inter-annotator reliability for human annotation tasks), but automated performance is typically well below that upper-bound level of performance.

/*We agree that imperfect performance of NLP models is an important limitation to their application to EHR data. We have updated the Strengths and Limitations section of the manuscript accordingly **(Page 14, Paragraph 4)**:

"However, a limitation of NLP models is that they can yield false positive and false negative instances that can introduce errors into secondary analyses of NLP-derived data from EHRs. NLP-derived data also depend on the presence of documented clinical information in free text records. The absence of documentation does not necessarily indicate the absence of a particular clinical construct and clinicians do not systematically document the absence of clinical features unless it is clinically relevant to do so."

Regarding the removal of personal health information (PHI), we have applied the Safe Harbor method (described in the *De-identification procedure* section of the main manuscript) to remove all structured data fields that could potentially contain PHI. To ensure data security and confidentiality, we do not provide access or enable the analysis of any unstructured/free text data within the NeuroBlu platform. Only structured data derived using NLP are made available for analysis. NLP is not employed to support the de-identification procedure and no unstructured data are made available for analysis to NeuroBlu users. We have updated the section on *Unstructured data* **(Page 9, Paragraph 2)** to clarify this point:

"To maintain data security and confidentiality, the original unstructured EHR data are not available for analysis. For social history/external stressors & MSE data, only the data derived using NLP for are available for analysis in the NeuroBlu platform."*/

In the absence of 100% human validation of the NLP results, therefore, in my opinion it's quite important to include an expected level of (in)correct classification along with automatically constructed variables -- not just in a published article (e.g. the results aggregating into eight categories, supplement page 16, are a start), but visibly for users in the platform whenever analysis results are being reported. An even better approach would be to provide users with confidence estimates at the per-item level, i.e. a confidence associated with each of the 241 NLP-derived binary features in every data record. One could even imagine the UI giving users control over how confidently an NLP-derived variable must be assigned in order to treat it as present in the record. (Cf. Jiang et al. 2006, "How does the system know it's right? Automated confidence assessment for compliant coding." In Perspect Health Inf Manag, Computer Assisted Coding Conference Proceedings.)

To be fair, the issue I'm raising here is present in many if not most user-facing NLP applications; for example, commercial sentiment platforms will frequently generate reports about percentages of positive, negative, and neutral statements in some dataset being analyzed as if one could be sure that those sentiment labels have all been assigned correctly. Still, I would like to think that clinical applications should be adhering to a higher standard, in much the same way that responsibly conducted political surveys report confidence intervals, rather than just giving point-estimate percentages.

/*We agree that providing numerical estimates of NLP accuracy to end users is crucial to ensure that data are analyzed in an appropriate way considering the potential for inaccurate classification (false positives, in particular) to bias results. We already provide full documentation of the NLP approach in the user documentation within the NeuroBlu platform including the precision statistics quoted in the supplementary material of the present manuscript. We have updated the Unstructured data section to clarify this **(Page 9, Paragraph 1)**:

"Information on NLP model development and precision statistics for MSE categories are also provided in the user guide on the NeuroBlu platform."*/

Finally, I also think it's important to think about this work in the context of broader developments in the analysis of medical data. The bar has been significantly lowered in recent years for applying rich machine-learning approaches in a wide variety of domains, but, as noted above, progress in the clinical space has been significantly hampered by lack of access to the same large, shared datasets by diverse teams. NeuroBlu represents a solid step in the right direction (to the extent it is actually available to a rich variety of researchers who can afford to use it) in terms of scale plus the ability to bring features communicated via text to the surface for statistics/analytics. A significant remaining obstacle to further progress, however, is the fact that a single, proprietary NLP approach is baked into the platform. And, in fact, it is an approach (LSTM with static word2vec embeddings; see Mukherjee et al. 2020) that has already been outstripped by more recent developments (contextual embeddings and the pretraining/finetuning approach using transformers). That doesn't make what's being done here wrong or inadequte, it's more of a missed opportunity (so far): it would be quite interesting to consider whether it would be possible to extend the platform's R/Python Code Engine capabilities to include a fuller range of machine learning approaches, including direct access to the (de-identified) text within the secure environment, so that users of the platform could themselves explore new classification approaches as the state of the art improves, and also expand to more NLP-derived features as motivated by the particular questions they are investigating.

/*Thank you for your supportive comments. We are fully committed to protecting data security and confidentiality. We adhere to and go beyond the minimum requirements set by data protection regulations and legislation. For the reasons described previously (pertaining to protection of PHI), we do not make any unstructured text available for analysis within the NeuroBlu platform itself and only structured data derived from the output of NLP models are available for analysis. However, we agree that more advanced NLP approaches could significantly improve the utility of derived data from unstructured EHR documents and we plan to improve upon our NLP methods in future work. We have updated the *Future plans* section of the manuscript to describe this **(Page 16, Paragraph 1)**:

"We plan to analyze these data to generate real-world evidence to better understand the factors related to poor mental health outcomes, to evaluate the impact of treatments, to develop more advanced NLP models with improved classification accuracy, and to develop predictive analytic tools that quantify these factors at individual patient level.[61]"*/

Reviewer: 1
Competing interests of Reviewer: None.

Reviewer: 2
Competing interests of Reviewer: I am on advisory boards with The Coleridge Initiative and NORC at the University of Chicago, both of which are nonprofit organizations whose activities include secure data enclaves.

VERSION 2 – REVIEW

REVIEWER	Resnick, Philip University of Maryland at College Park
REVIEW RETURNED	14-Feb-2022

GENERAL COMMENTS	Thanks for the careful attention and the work in revising the manuscript. In my review here I have made a recommendation or two that I think are important, but I'm going to recommend Accept, rather than potentially introducing another round of editing/re-review by recommending Minor Revision. I do hope you'll act on these suggestions, but they are not major enough to warrant slowing down the process of getting this paper out the door. I appreciate the authors' responses and I believe they have done a very good job of addressing my main areas of concern. I have just a few remaining notes. - The additional note on NLP error rates (page 14, line 50-51) does a reasonable job of addressing the issue I raised regarding imperfect NLP performance, but I think it's an important enough point that it also belongs under Article Summary, Strengths and Limitations. E.g. a bullet something like: "When used to extract information from clinical text, NLP models can yield false positives and false negatives. Therefore downstream analyses of NLP-derived data from EHRs need to take NLP error rates into account". - On a related note, in the update to Page 9, paragraph 1, I think "precision statistics" is not quite right for a clinical audience. (The relevant metrics are not just about precision, plus of course clinicians don't use that term, as the authors have addressed elsewhere by referring to positive predictive value.) I would prefer the term "error rates", since that's really what's at stake here, but if the authors are unhappy with introducing the term 'error' into the top-level discussion, then at least "accuracy statistics" or "performance statistics" would be a better. - I think the edits (Page 7, Paragraph 5) to describe the application of NLP specific to mental health EHR data for a clinical readership are generally well done. However, "NLP is a text mining technique" is inaccurate. I would suggest changing "NLP is a text mining technique that enables..." to "NLP is the sub-discipline of artificial intelligence that deals with naturally occurring human language, including techniques that enable...". In the following sentences, I would edit "involves" to "typically involves". Finally, I would change "evaluated on the previously annotated reference set" to "evaluated on a previously annotated, held-out reference set". - I appreciate the edits on Page 16, paragraph 1, regarding future plans to develop better NLP models using more advanced

	techniques. For what it's worth, I would strongly encourage the authors to consider saying they're planning to seek out collaborators to develop more advanced NLP models (as opposed to what's there now, which is strictly in-house). - Page 16 line 26 - typo: "with a academic partners"

VERSION 2 – AUTHOR RESPONSE

Response to reviewers - bmjopen-2021-057227.R1

Reviewer: 2
Prof. Philip Resnick, University of Maryland at College Park
Comments to the Author:
Thanks for the careful attention and the work in revising the manuscript. In my review here I have made a recommendation or two that I think are important, but I'm going to recommend Accept, rather than potentially introducing another round of editing/re-review by recommending Minor Revision. I do hope you'll act on these suggestions, but they are not major enough to warrant slowing down the process of getting this paper out the door.

Reviewer: 2
Competing interests of Reviewer: I am on advisory boards with The Coleridge Initiative and NORC at the University of Chicago, both of which are nonprofit organizations whose activities include secure data enclaves.

I appreciate the authors' responses and I believe they have done a very good job of addressing my main areas of concern. I have just a few remaining notes.

/*Thank you for your comments. We have responded to your recommendations below.*/

- The additional note on NLP error rates (page 14, line 50-51) does a reasonable job of addressing the issue I raised regarding imperfect NLP performance, but I think it's an important enough point that it also belongs under Article Summary, Strengths and Limitations. E.g. a bullet something like: "When used to extract information from clinical text, NLP models can yield false positives and false negatives. Therefore downstream analyses of NLP-derived data from EHRs need to take NLP error rates into account".

/*We agree this is an important limitation. We have appended this limitation to the penultimate bullet point in the **Strengths and Limitations** section as only a maximum of five bullet points are permitted.*/

- On a related note, in the update to Page 9, paragraph 1, I think "precision statistics" is not quite right for a clinical audience. (The relevant metrics are not just about precision, plus of course clinicians don't use that term, as the authors have addressed elsewhere by referring to positive predictive value.) I would prefer the term "error rates", since that's really what's at stake here, but if the authors are unhappy with introducing the term 'error' into the top-level discussion, then at least "accuracy statistics" or "performance statistics" would be a better.

/*We have changed “precision statistics” to “error rates” in **Page 9, paragraph 1**.*/

- I think the edits (Page 7, Paragraph 5) to describe the application of NLP specific to mental health EHR data for a clinical readership are generally well done. However, "NLP is a text mining technique" is inaccurate. I would suggest changing "NLP is a text mining technique that enables..." to "NLP is the sub-discipline of artificial intelligence that deals with naturally occurring human language, including techniques that enable...". In the following sentences, I would edit "involves" to "typically involves".

Finally, I would change "evaluated on the previously annotated reference set" to "evaluated on a previously annotated, held-out reference set".

/*We have made the suggested amendments in **Page 7, Paragraph 5**.*/

- I appreciate the edits on Page 16, paragraph 1, regarding future plans to develop better NLP models using more advanced techniques. For what it's worth, I would strongly encourage the authors to consider saying they're planning to seek out collaborators to develop more advanced NLP models (as opposed to what's there now, which is strictly in-house).

/*We are keen to seek out collaborations with NLP experts to further develop our NLP models and have updated **Page 16, Paragraph 1** accordingly.*/

- Page 16 line 26 - typo: "with a academic partners"

/*Thank you for highlighting this error in **Page 16, Paragraph 2**. We have corrected it.*/